

Extrakce sociálních sítí z byznys procesů

Mining Social Networks from Business Processes

Zadání bakalářské práce

Student:

Lukáš Panchártek

Studijní program:

B2647 Informační a komunikační technologie

Studijní obor:

2612R025 Informatika a výpočetní technika

Téma:

Extrakce sociálních sítí z byznys procesů
Mining Social Networks from Business Processes

Zásady pro vypracování:

Cílem práce je provedení průzkumu existujících přístupů, návrh a implementace vybrané nebo vlastní metody a aplikačního prostředí pro experimenty.

1. Průzkum a popis existujících přístupů.
2. Návrh a implementace vybrané nebo vlastní metody.
3. Návrh a implementace počítačové aplikace pro provádění experimentů.
4. Návrh, realizace a hodnocení experimentů.

Seznam doporučené odborné literatury:

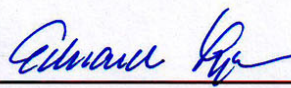
[1] Wil M. P. van der Aalst, Hajo A. Reijers, Minseok Song: Discovering Social Networks from Event Logs. Computer Supported Cooperative Work (CSCW) 14(6):549-593 (2005)
Další dle pokynů vedoucího bakalářské práce.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

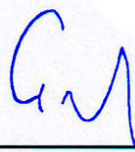
Vedoucí bakalářské práce: **Mgr. Miloš Kudělka, Ph.D.**

Datum zadání: 16.11.2012

Datum odevzdání: 07.05.2013



doc. Dr. Ing. Eduard Sojka
vedoucí katedry



prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě dne 30. dubna 2013



.....

Děkuji panu Mgr. Miloši Kudělkovi, Ph.D. za věcné připomínky a rady týkajících se dané problematiky a také blízkým osobám za vyjádřenou podporu.

Abstrakt

Cílem této bakalářské práce je popsat způsob, jak z daného protokolu událostí, který zachycuje události nějakého byznys procesu, vytvořit sociální síť a následně tuto síť analyzovat. Úkolem je popsat a vysvětlit techniky, které je možné pro tento problém využít. Následně jednu z možných metod implementovat, včetně návržení a implementace aplikace pro provádění experimentů.

Klíčová slova: extrakce, sociální síť, byznys proces, bakalářská práce, sociometrie, protokol událostí, analýza

Abstract

The aim of this bachelor thesis is to describe the way how we can use the event log (which captures the event of a business process) to create a social network and then analyze it. The task is to describe and explain the techniques that can be used for this problem. Then implement one of the possible methods, including the design and implementation of application for experimentation.

Keywords: mining, social network, business process, bachelor thesis, sociometry, event log, analysis

Seznam použitých zkratk a symbolů

ERP	– Enterprise Resource Planning
CRM	– Customer Relationship Management
XML	– Extensible Markup Language
XAML	– Extensible Application Markup Language

Obsah

1	Úvod	5
2	Přiblížení základních témat problému	6
2.1	Sociometrie	6
2.2	Byznys proces	6
2.3	Protokol událostí	6
3	Proces extrakce	9
3.1	Odhalování sociálních sítí	9
3.2	Problémy při odhalování sociálních sítí	12
4	Metriky	13
4.1	Metriky založené na (možné) kauzalitě	13
4.2	Metriky založené na společných případech	14
4.3	Metriky založené na společných aktivitách	15
4.4	Metriky založené na speciálních typech událostí	16
5	Analýza sociální sítě	17
5.1	Hustota (Density)	18
5.2	Prostřednictví (Betweenness Centrality)	19
5.3	Blízkost (Closeness Centrality)	20
5.4	Počet hran (Degree Centrality)	21
6	Implementace a praktické zpracování	23
6.1	Protokol událostí ve formě XML souboru	23
6.2	Zpracování dat	24
6.3	Zobrazení získaných výsledků	26
6.4	Popis experimentů	29
7	Závěr	32
8	Reference	33
	Přílohy	33
A	Příloha na CD/DVD	34

Seznam tabulek

1	Protokol událostí	9
2	Matice zachycuje kauzální závislosti mezi osobami	11
3	Matice aktivit jednotlivých osob	15
4	Hodnoty uzlů při použití metriky blízkosti	19
5	Nejkratší vzdálenosti mezi uzly pro hvězdicový graf	20
6	Nejkratší vzdálenosti mezi uzly pro řetězový graf	21
7	Hodnoty uzlů při použití metriky blízkosti	21

Seznam obrázků

1	Jednoduchý příklad sociogramu	7
2	Sociální síť	11
3	Sociální síť vytvořená na základě metriky založené na (možné) kauzalitě	14
4	Sociální síť vytvořená na základě metriky založené na společných případech	14
5	Hvězdicový a řetězový graf	18
6	Důležitost uzlů po použití metriky Prostřednictví (Betweenness Centrality)	20
7	Důležitost uzlů po použití metriky Blízkost (Closeness Centrality)	21
8	Důležitost uzlů po použití metriky Počet hran (Degree Centrality)	22
9	Uživatelská aplikace	26
10	Vysvětlivky pro uzly a hrany	27
11	Detail sítě, kde jsou hrany vykresleny bez ohledu na jejich váhu	28
12	Detail sítě, kde jsou hrany vykresleny s ohledem na jejich váhu	28
13	Zobrazení dat v uživatelské aplikaci	29
14	Vytvořená sociální síť s deseti uzly	30
15	Vytvořená sociální síť s padesáti uzly	31

Seznam výpisů zdrojového kódu

1	Ukázka struktury XML souboru	23
2	Funkce pro získání všech osob z protokolu událostí	25
3	Funkce pro získání všech případů dané osoby	25

1 Úvod

Tématem této bakalářské práce je popsat možné postupy pro získávání sociálních sítí z takzvaných byznys procesů a následně implementovat jeden z existujících postupů, vytvoření aplikace pro provádění experimentů a následný popis a zhodnocení získaných výsledků.

Existují situace, kdy je potřeba využít poznatky a získat různé údaje či data o nějaké skupině či společenství, ve kterém si jednotlivé osoby mezi sebou vytváří vztahy. Takto získané informace lze následně využít pro mnoho různých způsobů. Například pokud známe osobu, která má ze všech osob v sociální síti nejvíce propojení, lze o této osobě říci, že má v dané skupině největší slovo, a je proto nejlepší volbou pro oznámení nějaké informace ostatním. Praktickým případem, kdy bylo využito analýzy sociální sítě, bylo po teroristických útocích v roce 2001 ve Spojených státech amerických, kdy vyšetřovatelé vytvářeli a analyzovali teroristickou sociální síť a zjišťovali vztahy a spojení mezi jednotlivými teroristy. My se však zaměříme na sociální síť vytvářené z byznys procesů.

V kapitole 2 si vysvětlíme a popíšeme základní témata vztahující se k danému problému a u kterých je nutné vědět, co znamenají. V další kapitole 3 se podíváme na prvotní extrakci dat, která jsou potřeba pro vytvoření sociální sítě. Kapitola 4 obsahuje popis jednotlivých metrik, které můžeme využít na extrakci dat. Uvidíme, že ze stejných dat, které máme k dispozici, můžeme na základě vybrané metriky dostat zcela odlišnou sociální síť než při využití jiné metriky. V následující kapitole 5 je popsána analýza vzniklé sociální sítě, kdy už známe podobu sítě a můžeme vidět jednotlivé vazby a vztahy mezi těmito osobami. V poslední 6. kapitole je popsán praktický postup a náznaky implementace programu, který je schopen z daných dat sociální síť vytvořit.

2 Přiblížení základních témat problému

V této sekci si přiblížíme pojmy, které je nutná znát, abychom lépe pochopili celkovou problematiku. Jelikož během extrakce sociálních sítí z byznys procesů narazíme na různé pojmy, je dobré si celkový koncept rozdělit na části a tyto části poté vysvětlit jednotlivě.

2.1 Sociometrie

Jelikož výsledkem extrakce sociálních sítí z byznys procesů, jak je patrné, je sociální síť, zavedeme si pojem sociometrie. Sociometrie je teorie a metoda měření interpersonálních vztahů (sympatie a antipatie) v převážně menších sociálních skupinách. Snaží se tedy změřit a interpretovat kvalitu a kvantitu mezilidských vztahů. Zakladatelem sociometrie je rumunský lékař, sociolog a psychiatr Jakob Levy Moreno. Sociometrická metoda bývá nejčastěji aplikována v malé sociální skupině, jako je například pracovní skupina, třída, sportovní mužstvo. Pro naše potřeby je to hlavně skupina zainteresovaná v určitém byznys procesu. Sociometrie je nástroj, který pomáhá odhalit strukturu neformálních vztahů ve skupině.[1]

2.1.1 Sociogram

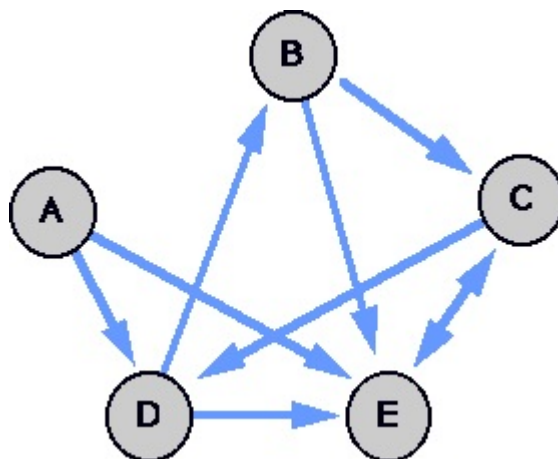
Velmi úzce je s pojmem sociometrie spjat pojem sociografie, která se stará o grafickou reprezentaci získaných výsledků, tedy grafické zobrazení pozice členů ve skupině a vzájemných vztahů mezi nimi. Grafické znázornění společenských vztahů se tedy nazývá sociogram. Sociogram je zpravidla vytvářen zpracováním matice vztahů mezi členy skupiny. Takto vzniklý sociogram je složen z uzlů, které reprezentují osoby či interprety zainteresované v daném společenství, naproti tomu hrany zobrazené v tomto grafu, reprezentují vztahy mezi osobami. Typický sociogram je zobrazen na 1 Další forma reprezentace získaných výsledků může být například také matice, ta se však nevyužívá tak často.[5][7]

2.2 Byznys proces

Jelikož úkolem je extrahovat sociální sítě z byznys procesů, definujeme si také pojem byznys proces. Byznys proces je po částech uspořádaná množina událostí, které společně realizují podnikový nebo strategický cíl, obvykle v kontextu organizační struktury definující funkce rolí a jejich vztahy. Tedy máme kolekci souvisejících a strukturovaných činností a úkolů, které provádí jednotlivé osoby a díky zpracování těchto činností se produkují specifické služby či produkty pro daného zákazníka.[3][4]

2.3 Protokol událostí

V této době již mnoho podnikových informačních systémů ukládá relevantní události v dostatečně strukturované podobě, díky které lze využít důležité informace k vytvoření sociální sítě. Například takzvané workflow řídicí systémy (tedy systémy s daným postupem jednotlivých navazujících událostí) si obvykle uchovávají informace o události,



Obrázek 1: Jednoduchý příklad sociogramu

respektive její začátek, informace kdo s touto událostí pracoval a také její dokončení. Existují samozřejmě také informace, které pro nás nemusí být podstatné a využitelné, těm se pochopitelně dále nevěnujeme.

Soubor obsahující tyto uložené informace se nazývá *protokol událostí*. Protokol událostí zachycuje jednotlivé události, tedy předávání práce v byznys procesu od jedné osoby k další a jejich aktivity v daných případech. Následným rozбором tohoto protokolu můžeme extrahovat sociální síť, která obsahuje všechny osoby podílejících se na jednotlivých událostech a v závislosti na použité metodě zachytit jejich vzájemné vztahy, ať už například podle společných aktivit těchto osob či stejných případů, na kterých se podílejí. [1]

2.3.1 Rozvržení protokolu událostí

V předchozím odstavci jsme si uvedly pojmy, které se ukládají v protokolu událostí. Pro lepší pochopení si podrobněji popíšeme, co jednotlivé pojmy znamenají. V protokolu událostí jsou podle času zaznamenány jednotlivé **události** (v odborných anglických textech se nejčastěji používá výraz *event*). Budeme předpokládat, že v naší situaci budeme potřebovat zachytit **aktivitu** (v angl. textech *activity*) a **případ** (v angl. textech *case*) jednotlivých **osob**. [1]

Případ (také může být označován jako instance procesu) je zjednodušeně řečeno věc, se kterou je zacházeno. Příkladem případu může být například objednávka zákazníka, pojistná událost, žádost o zaměstnání, atd.

Naproti tomu aktivita je libovolná operace nad případem, kde si jako příklad můžeme uvést „kontaktovat zákazníka“, „schválit objednávku“, atd.

Událost můžeme označit jako (p, a, o) , kde p značí případ, a představuje aktivitu a o

osobu, která v daném případě provádí aktivitu. Události jsou seřazeny v čase umožňující vyvodit kauzální závislosti mezi aktivitami a příslušné sociální interakce. Mějme událost (p, a_1, o_1) následovanou událostí (p, a_2, o_2) . Obě události odkazují na stejný případ p . Mezi těmito událostmi je však nějaké předání práce mezi osobou o_1 a osobou o_2 . Pokud se toto předání často opakuje, tedy že se předává práce z osoby o_1 na osobu o_2 , ale nikdy se nepředává práce z osoby o_1 na osobu o_3 , a to i přesto, že osoba o_3 má stejnou roli v organizaci jako osoba o_2 , pak z tohoto vyplývá, že vztah mezi osobou o_1 a osobou o_2 je silnější, než vztah mezi osobou o_1 a osobou o_3 . Díky těmto informacím, které dokážeme získat a zpracovat z protokolu událostí, je možné vytvářet sociální sítě vyjádřené pomocí sociogramu či matice. [1]

3 Proces extrakce

Cílem procesu extrakce je získat informace o procesech z protokolů událostí. Předpokládáme, že je možné zaznamenat události, kdy

- každá událost odkazuje na aktivitu (tj. dobře definovaný krok v procesu)
- každá událost odkazuje na případ
- každá událost odkazuje na osobu (tedy člověka vykonávající nebo zahajující aktivitu)
- události jsou totálně uspořádané

Jakýkoliv informační systém využívající transakčního systému jako je např. ERP, CRM či workflow řídicí systém poskytuje tyto informace v nějaké formě. Příklad protokolu událostí je zobrazen v tabulce 1.[2]

Případ	Aktivita	Osoba
A	1	Petr
B	1	Petr
C	1	Nikol
C	2	Monika
A	2	Jakub
A	3	Petr
B	3	Jakub
D	1	Nikol
B	2	Petr
B	4	Jan
E	1	Nikol
D	3	Monika
A	4	Jan
C	3	Nikol
C	4	Jan
D	2	Nikol
E	4	Andrea
E	5	Andrea
D	3	Jan

Tabulka 1: Protokol událostí

3.1 Odhalování sociálních sítí

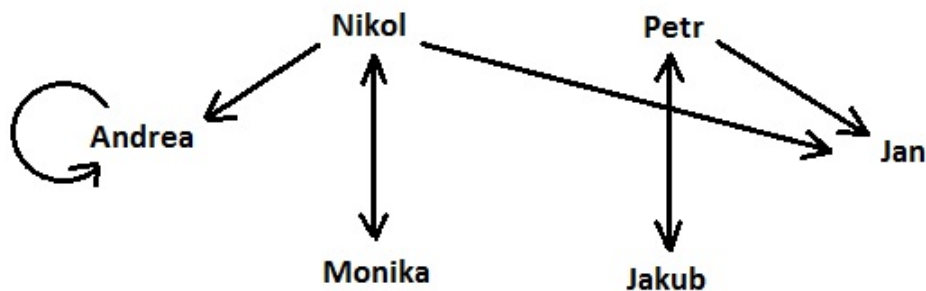
Při extrakci modelu procesu z protokolu událostí se zaměřujeme na různorodé aktivity procesu a jeho závislosti. Při odvozování jednotlivých rolí a dalších organizačních entit je kladen důraz na vztahy mezi osobami nebo skupinami a procesy. Další variantou

je zaměřit se na vztahy mezi jednotlivci či skupinami jednotlivců, kteří jsou v procesu angažování.[1]

Pokud se podíváme na protokol událostí, který je jako příklad uveden v tabulce 1, můžeme z něj mimo jiné vyvodit, že přestože Monika a Jakub provádějí stejné aktivity (konkrétně aktivity 2 a 3) a tím pádem mají tyto dvě osoby s největší pravděpodobností stejnou roli (jak jsme si uvedli dříve, aktivita je nějaká operace nad případem, tedy pokud dvě osoby mají stejné aktivity, provádějí stejné operace), přesto však nejsou angažované ve stejných případech. Jakub vždy pracuje s Petrem (případy A a B) a Monika pracuje vždy s Nikol (případy C a D). Jakub a Monika nikdy nepracují spolu na stejném případě a vypadá to, že ani Petr nikdy nepracuje s Monikou, a Nikol nikdy nepracuje s Jakubem. Samozřejmě, že protokol událostí z tabulky 1 má velmi malý počet vzorků se kterými můžeme pracovat a vyvozovat tak jednotlivé vztahy mezi zúčastněnými osobami. Avšak můžeme jednoduše předvést, jak odvodit z protokolu událostí vztahy mezi osobami a následně získat sociální síť ve formě sociogramu.[1]

Ukážeme si, jak by mohl vypadat sociogram vytvořený z protokolu událostí na základě předání práce z jedné osoby na druhou. Z obrázku 2 je patrné, že každý uzel představuje jednu konkrétní osobu angažovanou v protokolu událostí z tabulky 1. Jelikož v protokolu událostí je uvedeno šest konkrétních osob, sociogram uvedený v obrázku 2 bude obsahovat šest uzlů. Každá hrana mezi uzly představuje předání práce z jedné osoby na druhou. Definujme si, že hrana se vykreslí mezi uzly X a Y právě tehdy, pokud po aktivitě prováděné X bezprostředně následuje aktivita, kterou provádí Y, a to ve stejném případě. Tedy práce se předává z X do Y. Konkrétněji se podívejme do našeho protokolu událostí, kde v případě A a v případě B se práce předává od Petra k Jakubovi. Avšak, jak je ze sociogramu uvedeném na obrázku 2 patrné, je zde předávání práce i naopak – tedy od Jakuba k Petrovi. Jediná Andrea předává práci sobě samotné (konkrétně tak jedná v případě E, kdy prvně provede aktivitu 4 a následně aktivitu 5) a Jan nepředává práci nikomu, tudíž můžeme s jistotou říci, že Jan je poslední osobou, která provádí aktivitu v daném případě a ukončuje jej, konkrétně se jedná o případ D.

Samozřejmě, jelikož máme velmi malý vzorek zaznamenaných událostí, zdá se tento sociogram i jednotlivé vztahy triviální, avšak při větším množství záznamů dostáváme složitější sociogramy s velkým množstvím propojení jednotlivých uzlů, vytvářející zajímavé struktury.



Obrázek 2: Sociální síť

Existuje několik způsobů, jak mohou být sociogramy zobrazovány. Sociogramy mohou být dle potřeby zobrazeny buď jako neorientované či orientované grafy (tedy od koho ke komu je hrana směřována). Dále jednotlivé vazby, tedy hrany mezi uzly, mohou mít váhu či nikoliv. Z pohledu sociologie může mít tato váha informaci, zda se jedná o silný či slabý vztah mezi dvěma uzly, respektive mezi dvěma osobami.[1]

Jak jsme si uvedli, sociogram není jediný způsob, jak zobrazit výsledné vztahy jednotlivých osob podílejících se svými aktivitami v konkrétních případech. Tyto vztahy můžeme zobrazit také pomocí matice. Výsledná matice našeho protokolu událostí zobrazeného v tabulce 1 je zobrazena v tabulce 2. Levý sloupec obsahuje osoby, které práci předávají. Horní řádek popisuje naopak osoby, kterým je práce předávána. Tato matice popisuje kauzální závislosti jednotlivých osob, tudíž je zde možnost předat práci sebe sama, jak můžeme vidět u Andreji. Také se zde počítají jednotlivé předání práce od jedné osoby k druhé. Tyto hodnoty se dají využít k určení síly vztahů jednotlivých osob mezi sebou. Čím větší číslo, tím mají osoby k sobě blíže, jelikož spolu častěji pracují. Samozřejmě u jiných metrik, které je možno použít pro získávání sociálních sítí z byznys procesů, bude matice vypadat jinak a obsahovat jiné hodnoty.[1]

	Petr	Nikol	Monika	Jakub	Jan	Andrea
Petr	0	0	0	2	2	0
Nikol	0	0	2	0	2	1
Monika	0	2	0	0	0	0
Jakub	2	0	0	0	0	0
Jan	0	0	0	0	0	0
Andrea	0	0	0	0	0	1

Tabulka 2: Matice zachycuje kauzální závislosti mezi osobami

Z takovéto matice, obsahující vztahy mezi jednotlivými osobami, není složité vytvořit sociální síť ve formě sociogramu, který je přehlednější a lze z něj jednotlivé vztahy lépe vidět a pochopit.

3.2 Problémy při odhalování sociálních sítí

Pro realističtější situace existuje několik komplikujících faktorů, které mohou nastat při velkém množství zaznamenaných informací, tedy při velkém počtu zachycených událostí v protokolu událostí. V následujících podkapitolách si tyto problémy popíšeme.

3.2.1 Kompletnost

Pro větší workflow systémy a modely, které mohou obsahovat alternativy a paralelní zpracování, neobsahují jejich protokoly událostí obvykle všechny možné propojení. Například, pokud budeme mít 10 aktivit, které mohou být provedeny paralelně, celkový počet prokládání bude $10! = 3628800$. Proto není možné, aby každé prokládání bylo v protokolu událostí přítomno. Navíc některé cesty v modelu procesu mají nízkou pravděpodobnost existence, a proto zůstanou neodhaleny. Podobné situace mohou nastat také pro sociální sítě. Například osoba má roli, ale jen shodou okolností neprovedla některé nebo všechny činnosti odpovídající této roli. Další příkladem může být, pokud dvě osoby pracují často společně, ale v době, kdy se shromažďovala data, byla jedna osoba mimo dění a tím pádem se nezaznamenaly její aktivity. V důsledku těchto možných situací není protokol úplný v tom smyslu, že nezachycuje možné či typické chování.[1]

3.2.2 Šum

V protokolu událostí mohou být některé z jeho části nepřesné, neúplné či odkazují na nějaké výjimky. Události mohou být zaznamenány špatně, kvůli lidským či technickým chybám. Události mohou v protokolu chybět, pokud některé z aktivit jsou zadávány ručně nebo jsou zpracovány jiným systémem. Události mohou také odkazovat na vzácné nebo nežádoucí události. Vezměme si například workflow řídicí systém v nemocnici. Pokud je v důsledku časového tlaku pořadí dvou událostí (např. událost udělat rentgen a událost odstranit odtok) prohozeno, neznamená to, že by to bylo součástí regulérního lékařského protokolu. Také mohou nastat dvě kauzálně nesouvisející události (např. vzít vzorek krve a smrt pacienta), aniž by to znamenalo příčinný vztah (tj. odběrem krve nemělo vést ke smrti pacienta, byla to naprostá náhoda). Je zřejmé, že výjimky, které jsou zaznamenány pouze jednou, se nesmí automaticky stát součástí regulérního workflow.[1]

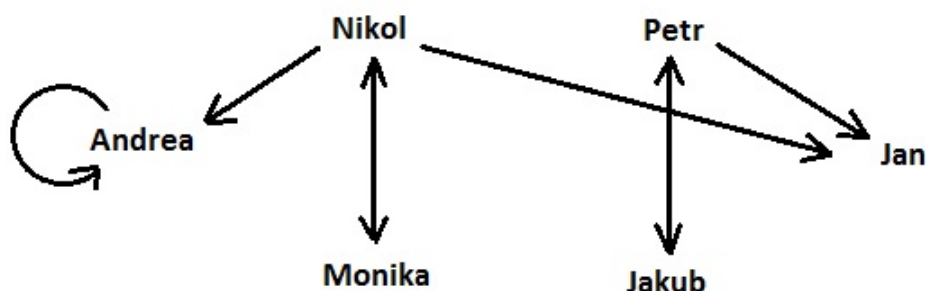
4 Metriky

Pomocí metrik můžeme z daného souboru, který obsahuje uložené informace o jednotlivých událostech (protokol událostí), získat sociální síť. Existuje několik různých způsobů a metod, jak daný protokol událostí zpracovat. Výsledný sociogram také závisí na zvolené metodě, čili při zvolení jedné metody pro zpracování protokolu událostí nám vznikne jiná sociální síť než při zvolení druhé. Jednotlivé metriky, které je možné použít pro extrakci sociálních sítí z byznys procesů, si popíšeme v následujících podkapitolách. U každé metriky si také uvedeme sociogramy, které by vznikl jejím použitím. Využijeme přitom protokol událostí zobrazen v tabulce 1.

4.1 Metriky založené na (možné) kauzalitě

Metriky založené na (možné) kauzalitě monitorují pro jednotlivé případy, jak se práce přesouvá mezi osobami, které jsou součástí těchto případů. Jedním z příkladů takové metriky je *předávání práce* (handover of work). Tedy že nastane situace, že je práce předána od osoby A k osobě B , kde existují dvě po sobě jdoucí aktivity, kdy první aktivita je dokončena osobou A a druhá aktivita dokončena osobou B . Samozřejmě, že v rámci jednoho případu. Toto pojetí může být modifikováno různými způsoby. Například znalost procesu struktury může být použita ke zjištění, zda je mezi oběma aktivitami opravdu kauzální závislost. Je také možné brát v úvahu nejen přímé sledy aktivit, ale i nepřímá posloupnost aktivit pomocí tzn. kauzálního pádového faktoru β (causality fall factor). Pokud máme tři aktivity mezi činnostmi, která byla dokončena osobou i a činností zakončenou osobou j bude kauzální pádový faktor β^3 . Související metrika využívá „subdodavatele“. Hlavní myšlenkou je spočítat kolikrát jedinec j vykonával aktivitu mezi dvěma aktivitami, které vykonával jedinec i . To může indikovat, že proběhla subdodávka práce z jedince i do jedince j . [1]

Na následujícím obrázku 3 je zobrazena sociální síť na základě předání práce z jedné osoby na druhou. Bereme v potaz přímé předání práce z osoby na osobu. Nikol předává práci Andree (případ E), proto vede vztah od uzlu s názvem Nikol k uzlu s názvem Andree, atd. Vazby v této sociální síti by mohly mít i svou váhu, tedy jak silný je vztah mezi uzly (osobami), která by byla daná počtem předávek práce od výchozí osoby k druhé osobě.

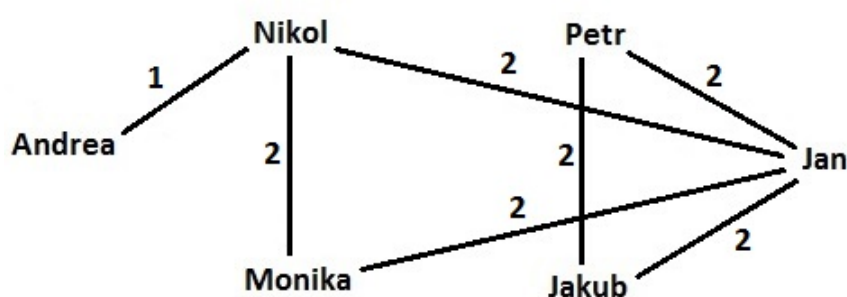


Obrázek 3: Sociální síť vytvořená na základě metriky založené na (možné) kauzalitě

4.2 Metriky založené na společných případech

Metriky založené na společných případech ignorují kauzální závislosti, ale jednoduše počítají, jak často dvě osoby vykonávají aktivity pro stejný případ. Pokud jednotlivci spolupracují na mnoha případech, budou mít silnější vztah než jednotlivci zřídka pracující společně.[1]

Jak můžeme vidět na obrázku 4, sociální síť vytvořená z protokolu událostí zobrazeného v tabulce 1 pomocí této metriky nemá směrované vazby mezi jednotlivými uzly, jako v případě použití metriky založené na (možné) kauzalitě. Je to z toho důvodu, že osoby angažované v jednom případě, mají vždy obousměrný vztah mezi sebou. Není možné, aby od jedné osoby šel vztah ke druhé, a z druhé nešel k té první. Z obrázku 4 je také patrné, že jednotlivé vazby mají svou váhu, konkrétně zobrazují počet případů, ve kterých jsou dané osoby společně zainteresované. Z ukázkové sociální sítě je také patrné, že Jan je angažován v nejvíce případech (má největší počet spojení) a Nikol má silnější vztah s Monikou a Janem než s Andreou, jelikož má s nimi více společných případů.



Obrázek 4: Sociální síť vytvořená na základě metriky založené na společných případech

4.3 Metriky založené na společných aktivitách

Metriky založené na společných aktivitách neberou v potaz, zda jednotlivci pracují společně na společných případech, ale zaměřují se na aktivity, které dané osoby vykonávají. Vychází z toho, že osoby dělající podobné věci mají silnější vztahy než lidé dělající úplně odlišné věci. Každý jedinec má „profil“ založený na tom, jak často provádí konkrétní činnosti. Existuje mnoho způsobů, jak měřit vzdálenosti mezi dvěma osobami.[1]

Následující matice zobrazuje pro jednotlivé osoby z protokolu událostí z tabulky 1, kolikrát vykonávají jednotlivé aktivity.

	aktivita 1	aktivita 2	aktivita 3	aktivita 4	aktivita 5
Petr	2	1	1	0	0
Nikol	3	1	1	0	0
Monika	0	1	1	0	0
Jakub	0	1	1	0	0
Jan	0	0	1	3	0
Andrea	0	0	0	1	1

Tabulka 3: Matice aktivit jednotlivých osob

Z této matice aktivit můžeme pomocí různých metod změřit vzdálenost mezi jednotlivými účastníky. Mezi metody, které se na výpočet vzdáleností využívají, patří mimo jiné také:

- Minkovského vzdálenost
- Hammingova vzdálenost
- Pearsonův korelační koeficient

Každá z těchto metrik využívá jiných postupů, jak vypočítat vzdálenost mezi dvěma osobami, vždy však pracuje s příslušným řádkovým vektorem. Rozeberme si například metriku Minkovského vzdálenost. Tato metrika vypočítá vzdálenost mezi osobami jako součet absolutních rozdílů jednotlivých prvků vektorů. Vezměme si například z tabulky 3 Nikol a Moniku, pokud má Nikol vektor aktivit (3, 1, 1, 0, 0) a Monika (0, 1, 1, 0, 0) vzdálenost bude pomocí Minkovského metriky vypočítána jako:

$$D_M = (|3 - 0|^n + |1 - 1|^n + |1 - 1|^n + |0 - 0|^n + |0 - 0|^n)^{1/n} = 3, \text{ kde } n = 1.$$

Při využití metriky Hammingova vzdálenost se využívá porovnávání dvou řetězců, v našem případě vektorů. Metrika počítá počet pozic, na kterém se dva vektory liší (v našem případě se liší právě tehdy, když jedna hodnota bude rovna nule a druhá bude jakékoli kladné číslo). Čím nižší je výsledné číslo, tím blíže k sobě dané osoby mají. Vezměme si například Nikol a Moniku, pokud má Nikol vektor aktivit (3, 1, 1, 0, 0) a Monika (0, 1, 1, 0, 0) potom Hammingova vzdálenost mezi nimi bude rovna 3. Vypočítáme-li Hammingovu vzdálenost mezi Monikou a Jakubem, zjistíme, že ta je rovna 0 a tudíž můžeme říci, že

Monika má silnější vztah s Jakubem než s Nikolou.

V případě použití metriky Pearsonův korelační koeficient může výsledek nabývat hodnot v intervalu $< -1, 1 >$. Pokud korelace dosáhne hodnoty $+1$ znamená to, že vzdálenost mezi měřenými osobami je malá a tyto osoby mají mezi sebou opravdu silný vztah. Naopak pokud se korelace blíží hodnotě -1 vzdálenost roste a vztah mezi osobami je slabý.[1] Pearsonův korelační koeficient se vypočítá pomocí vzorce:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Vzdálenost mezi Monikou a Nikol bude $r = 0$, tedy v půlce intervalu. Oproti tomu vzdálenost mezi Monikou a Jakubem bude $r = 1$, tedy maximální možná hodnota a vztah je nejsilnější.

4.4 Metriky založené na speciálních typech událostí

Metriky založené na speciálních typech událostí berou na rozdíl od předchozích metrik v úvahu typ události. Dosud jsme předpokládali, že událost obsahovala dokončení aktivity. Nicméně existují také události, které přerozdělují činnost z jedné osoby na druhou. Například pokud osoba A často deleguje práci na osobu B, ale ne naopak, pak je pravděpodobné, že osoba A je v hierarchickém vztahu s osobou B. Z pohledu analýzy sociální sítě jsou tyto postřehy obzvláště zajímavé, protože představují přímé mocenské vztahy.[1]

5 Analýza sociální sítě

Jakmile získáme z aplikace, která dokáže zpracovat protokol událostí, sociální síť, je možné z daného sociogramu analyzovat mnoho informací a údajů o skupině i jednotlivých účastnících vytvářející danou skupinu. Prvně si musíme uvědomit, jaký přístup k dané sociální síti zvolíme. Existují, respektive používají se dva typy přístupů, sociocentrický či egocentrický. Sociocentrický přístup považuje skupinu jako celek (centrem dění se stává celá skupina), naproti tomu egocentrický přístup považuje skupinu jako síť jednotlivců, kdy např. můžeme analyzovat vztahy mezi přáteli dané osoby. V mnoha věcech jsou tyto přístupy podobné, v obou případech je vstupním bodem pro analýzu graf, ve kterém uzly představují osoby a hrany mezi jednotlivými uzly představují vztahy. Jak jsme si dříve uvedli, graf může být neorientovaný či orientovaný (např. z bodu A vede šipka do bodu B, ale ne naopak). Jednotlivé vztahy mohou být označeny binárně (buď existuje nebo ne) nebo váženě („+“ nebo „-“, reálné číslo). Hodnota váhy daného vztahu udává, jak je daný vztah silný a dá se následně také využít.[1]

Z analýzy výsledné sociální sítě se můžeme dozvědět mnoho zajímavých faktů, jako např. „sociální vzdálenosti“ mezi jednotlivými uzly respektive osobami v sociální síti, které mohou být nejen v krátké, ale také v dlouhé vzdálenosti. Celkovou hustotu dané sociální sítě. Nalézt jednotlivce, kteří jsou „izolovaní“ od všech ostatních (tedy nemají žádné vztahy), nalézt tzn. „středky“ či „spojky“, které naopak mají mnoho propojení v sociální síti. Můžeme také vysledovat méně patrné pojmy, které se k analýze vážou. Například jedinec, který je propojen s lidmi s mnoha spojeními, je pokládán za více významný uzel v síti, než jedinec, který má mnoho spojení s lidmi, kteří mají málo propojení.

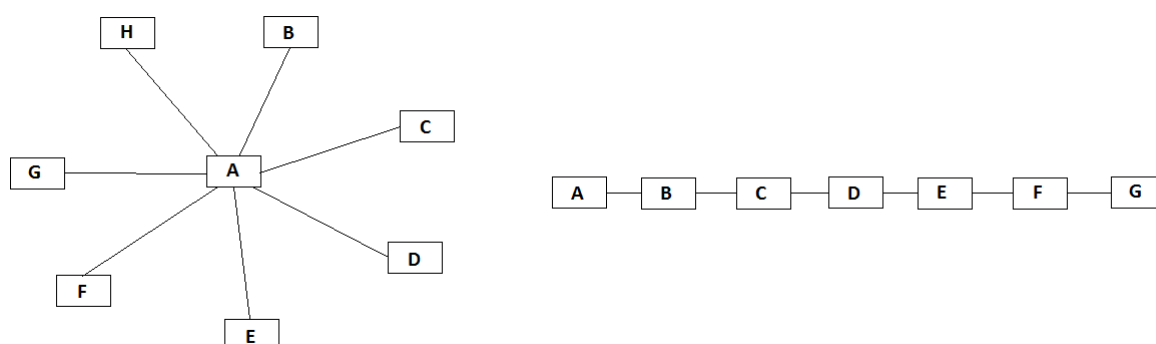
Když se na sociogram podíváme z matematického hlediska, dostaneme graf (P, R) , kde P je soubor jedinců (v rámci protokolu událostí se jedná o osoby) a $R \subseteq P \times P$ (R je podmnožina množiny kartézského součinu P a P). Pokud je graf neorientovaný, podmnožina R je symetrická. Pokud je graf vážený, existuje doplňková funkce W , která přiřadí hodnotu všem prvkům v podmnožině R . [1] [1]

Jak jsme si uvedli, jednotlivé uzly v grafu jsou pospojovány s ostatními uzly pomocí hran, které představují vztahy. Pro metriky, které si popíšeme níže, budeme potřebovat znát pojem geodetická vzdálenost. Geodetická vzdálenost dvou uzlů není nic jiného, než nalezení nejkratší cesty v grafu mezi těmito dvěma uzly. U vážených grafů bereme v potaz váhu jednotlivých spojení respektive hran, naopak u nevážených grafů, kde hrany nenesou žádnou hodnotu, samozřejmě v potaz nic brát nemůžeme, ale naopak hledáme nejkratší cestu tvořenou co nejmenším počtem hran. U vážených grafů není podmínkou, že cesta od jednoho uzlu ke druhému tvořená nejmenším počtem hran, je ta nejkratší, jelikož jednotlivé hrany mají svou váhu. [1]

Pokud se při analýze grafu zaměříme na jednotlivé osoby, respektive uzly, můžeme definovat mnoho zajímavých pojmů, které se k tomuto uzlu vážou. Pokud jsou všechny uzly v relativně krátké vzdálenosti od daného vybraného uzlu a jsou vzdáleny od tohoto uzlu

geodetickou vzdáleností, můžeme tento uzel nazvat centrálním uzlem. Existuje několik rozdílných metrik, které slouží pro výpočet centrálnosti uzlu. V následujících podkapitolách tyto metriky vysvětlíme a popíšeme, jakým způsobem vypočítávají centrálnost uzlu. Centrálností uzlu myslíme středovost tohoto uzlu, tedy čím více je uzel středový, tím více se stává důležitějším v celé sociální síti.[8]

Abychom jednotlivé metriky ještě lépe pochopili, ukážeme si na následujících jednoduchých grafech, jak se mění centrálnost jednotlivých uzlů při použití daných metrik.



Obrázek 5: Hvězdicový a řetězový graf

5.1 Hustota (Density)

Ještě před popisem jednotlivých metrik centrálnosti uzlu si vysvětlíme pojem hustota sociální sítě. Hustota sociální sítě určuje, jak moc jsou uzly v dané sociální síti propojené mezi sebou, čili bere v potaz maximální počet spojení v daném grafu a aktuální počet spojení v tomto grafu.[1]

Definujme si hustotu grafu jako počet prvků v R děleno maximální počet prvků, např. v orientovaném grafu existuje n^2 možných propojení (včetně vlastních smyček), kde n je počet uzlů. Potom tedy hustota grafu znázorněného na obrázku 2 bude $8/(6 \cdot 6) = 0,22$. [1]

Samozřejmě existují i grafy, kde neexistuje propojení uzlu na sebe sama (vlastní smyčka), v tomto případě každý uzel má teoretický počet propojení $n - 1$. To znamená, že maximální počet spojení v celé sociální síti není n^2 , ale $n \cdot (n - 1)$.

Pokud se podíváme na obrázek 5 a předpokládáme, že grafy jsou neorientované (tedy pokud existuje vztah od uzlu i k uzlu k , potom existuje také vztah od uzlu k k uzlu i) a není zde možnost vlastních smyček, bude u hvězdicového grafu hustota počítána jako $14/(8 \cdot 7) = 0,25$. U řetězového grafu bude tato hustota $12/(7 \cdot 6) = 0,286$.

5.2 Prostřednictví (Betweenness Centrality)

Tato metrika počítá počet nejkratších cest mezi uzlem i a uzlem j , na které se nachází uzel k , u kterého počítáme centrálnost pomocí prostřednictví. Uzel k má tedy možnost ovlivňovat tok informací mezi uzly i a j a umožňuje dobrou viditelnost všeho, co se v síti děje. Tedy uzel, který má nejvyšší hodnotu centrálnosti pomocí této metriky, je nejčastější uzel, který leží na nejkratších cestách mezi jinými dvěma uzly.[1]

Vezměme si zadaný hvězdicový graf. Spočítáme si celkový počet průchodů v tomto grafu. Z uzlu A vede ke všem ostatním uzlům přímá cesta, kdy se neprochází žádným dalším uzlem, čili počet průchodů je zatím roven nule. Přejdeme k uzlu B . Z tohoto uzlu vede celkem šest průchodů (vždy se prochází přes uzel A). Podobně jako uzel B , i z ostatních uzlů vede šest průchodů, čili celkový počet průchodů v hvězdicovém grafu je 42.

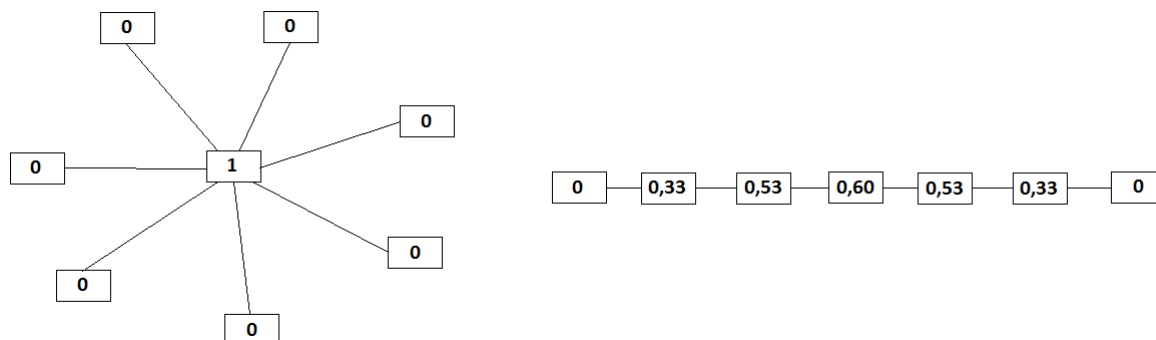
Nyní potřebujeme znát ke každému uzlu, kolikrát je součástí nějakého průchodu, respektive kolikrát se daný uzel vyskytuje v nejkratší cestě mezi jinými dvěma uzly. Z hvězdicového grafu je patrné, že jediný uzel, který může ležet v nejkratší cestě mezi dalšími uzly, je uzel A . Ostatní uzly se nevyskytují v žádné cestě mezi dalšími uzly, čili jejich hodnota bude nulová. Uzel A se vyskytuje celkem ve 42 nejkratších cestách, jelikož krajní uzly vždy prochází přes uzel A , aby se dostaly k dalšímu krajnímu uzlu.

Pokud známe počet nejkratších cest, které prochází jednotlivými uzly, můžeme spočítat těmto uzlům jejich hodnotu prostřednictví v sociální síti. Ta se spočítá jako: počet nejkratších cest, ve kterých uzel leží děleno celkový počet nejkratších cest v grafu.

Obdobně by se spočítaly hodnoty pro uzly řetězového grafu. Jen při pohledu můžeme vidět, že krajní uzly A a G budou mít nulovou hodnotu, jelikož leží na krajích řetězového grafu, a tudíž ani jeden z nich nemůže ležet v jakékoliv nejkratší cestě dvou uzlů. V následující tabulce 4 jsou vypočítány hodnoty pro všechny uzly jak v hvězdicovém, tak v řetězovém grafu.[11]

uzel	hvězdicový graf	řetězový graf
A	1	0
B	0	0,33
C	0	0,53
D	0	0,60
E	0	0,53
F	0	0,33
G	0	0
H	0	

Tabulka 4: Hodnoty uzlů při použití metriky blízkosti



Obrázek 6: Důležitost uzlů po použití metriky Prostřednictví (Betweenness Centrality)

5.3 Blížkost (Closeness Centrality)

Jak je z názvu patrné, tato metrika využívá vzdálenosti daného uzlu k ostatním. Tedy uzel je považován za důležitý právě tehdy, pokud je tento uzel v relativně krátké vzdálenosti od všech ostatních uzlů v síti. Uzly s vysokou mírou blízkosti mají velký vliv na to, co se v dané sociální síti odehrává.[1][11]

Výpočet této metody spočívá v sečtení všech nejkratších vzdáleností od uzlů v síti k ostatním uzlům a následně se takto získaný součet umocní na -1.

	A	B	C	D	E	F	G	H
A	0	1	1	1	1	1	1	1
B	1	0	2	2	2	2	2	2
C	1	2	0	2	2	2	2	2
D	1	2	2	0	2	2	2	2
E	1	2	2	2	0	2	2	2
F	1	2	2	2	2	0	2	2
G	1	2	2	2	2	2	0	2
H	1	2	2	2	2	2	2	0

Tabulka 5: Nejkratší vzdálenosti mezi uzly pro hvězdicový graf

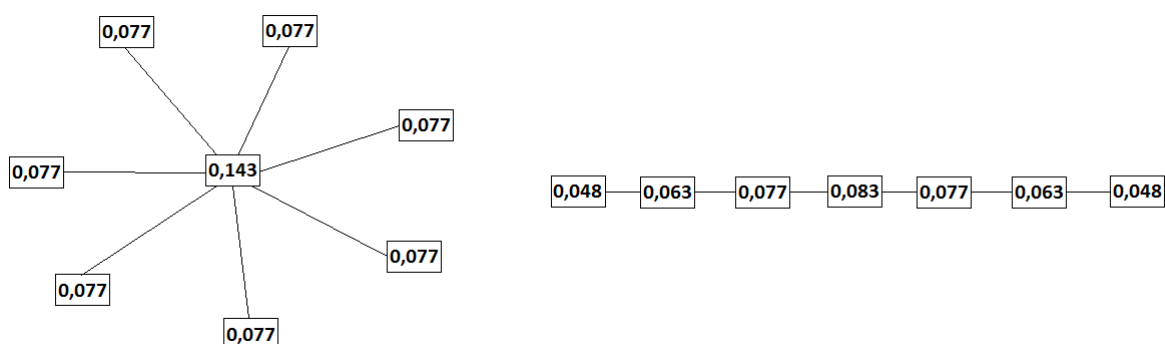
Z údajů zobrazených v tabulkách 6 a 5 následně můžeme určit hodnoty centrality jednotlivých uzlů, tedy sečtením všech nejkratších vzdáleností daného uzlu a následným umocněním tohoto součtu hodnotou -1. Jednotlivé hodnoty uzlů jsou zobrazeny tabulce 7.

	A	B	C	D	E	F	G
A	0	1	2	3	4	5	6
B	1	0	1	2	3	4	5
C	2	1	0	1	2	3	4
D	3	2	1	0	1	2	3
E	4	3	2	1	0	1	2
F	5	4	3	2	1	0	1
G	6	5	4	3	2	1	0

Tabulka 6: Nejkratší vzdálenosti mezi uzly pro řetězový graf

uzel	hvězdicový graf	řetězový graf
A	0,143	0,048
B	0,077	0,063
C	0,077	0,077
D	0,077	0,083
E	0,077	0,077
F	0,077	0,063
G	0,077	0,048
H	0,077	

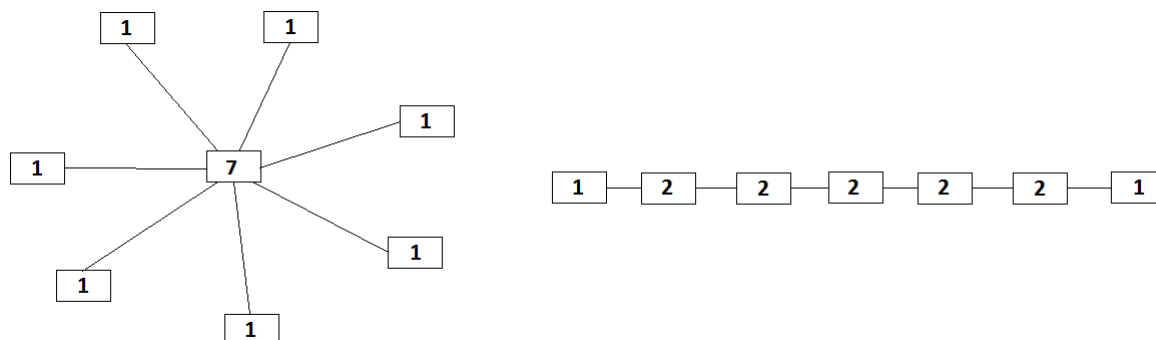
Tabulka 7: Hodnoty uzlů při použití metriky blízkosti



Obrázek 7: Důležitost uzlů po použití metriky Blízkost (Closeness Centrality)

5.4 Počet hran (Degree Centrality)

Tato metrika pro výpočet centrálnosti uzlu je z daných metod nejvíce intuitivní a také velmi jednoduchá. Při počítání středovosti uzlu touto metodou se využívá pouze počet přímých vazeb k dalším uzlům. Měří tedy aktivitu daného uzlu. Uzly s vysokou hodnotou počtu hran jsou označovány jako „spojky“ nebo „střed“ dané sítě a jsou mnohem důležitější než uzly s nízkou hodnotou. Jednoduše řečeno, uzel s největším počtem spojení je nejdůležitější v dané sociální síti.[6]



Obrázek 8: Důležitost uzlů po použití metriky Počet hran (Degree Centrality)

Jak můžeme z výše daných grafů zobrazených na obrázcích 6, 7, 8 vidět, každá z metrik počítá středovost uzlu zcela jinak a tím pádem se jednotlivé hodnoty uzlů liší. Proto je před analýzou sociální sítě důležité si uvědomit, co nás konkrétně na dané síti zajímá. Zda nám jde o to, který uzel má nejvíce spojení nebo nám záleží na tom, který uzel je nejbližší všem ostatním uzlům či nám jde o zcela něco jiného.

Další pojmy a metody, které souvisejí s analýzou sociální sítě a mohou být v tomto směru využívány, jsou **vyzařování** (z anglického pojmu emission) **zdroje**, které se počítá jako součet vazeb vycházejících z daného uzlu, tj. $\sum_j W_{i,j}$. **Příjem** (reception) **zdroje**, kdy se počítá počet vazeb vcházejících do uzlu, tj. $\sum_j W_{j,i}$. Další metrikou může být **stanovení stupně** (determination degree), kdy se od přijímaných vazeb odčítají odchozí vazby, tj. $\sum_j W_{j,i} - W_{i,j}$. Jinou zajímavou metodou pro analýzu je také **sociometrický status** (sociometric status), který je určen pomocí součtu vstupních a výstupních vztahů a určuje pozici dané osoby v sociální síti.[1]

Použitím těchto výše uvedených metrik a grafického zobrazení sociální sítě, můžeme analyzovat různé aspekty dané sociální sítě a nalézt celou řadu zajímavých informací. Například nalezení oblastí s hustým propojením či naopak jedince, kteří se do dané sociální struktury příliš nezapojují.

6 Implementace a praktické zpracování

Tato kapitola bude popisovat praktický přístup k dané problematice a popisovat řešení jednotlivých postupů vedoucí k zobrazení sociální sítě z předem daného protokolu událostí uloženého jako XML soubor. Popisuje také postupy analýzy takto vzniklé sociální sítě.

Implementace aplikace pro zpracování dat a jejich zobrazení je realizována v jazyce C#. Uživatelské rozhraní je vytvořeno s využitím technologie WPF (Windows Presentation Foundation), která využívá pro tvorbu uživatelského prostředí značkovací jazyk XAML. Výsledná struktura programu obsahuje několik tříd, kde každá třída má své využití. Pro získávání potřebných dat z XML souboru jsou vytvořeny třídy `ParseXML` a `Mapper_node`, které jako jediné zpracovávají datový XML soubor. Aplikační vrstva se tedy vůbec nezaobírá dolování dat ze souboru, ale pouze daná data dále zpracovává k zobrazení a potřebná data také využívá na analýzu sítě.

Grafické prostředí je tedy vytvořeno pomocí značkovacího jazyka XAML a jeho kód na pozadí využívá dále potřebné třídy `ParseXML`, `Node` (ta získává hodnoty pomocí třídy `Mapper_node`), `UserInfo` a hlavně třídu `MainViewModel`, která se stará o data nutná k zobrazení. `MainViewModel` využívá tříd `NodeG` pro definici jednotlivých uzlů, `EdgeG` pro definici vztahů mezi uzly a také třídu `GraphG`.

Pro zobrazení sociální sítě je využito volně dostupného frameworku *GraphSharp*. Tento framework dokáže velmi kvalitně vykreslit sociální síť za pomoci mnoha různých algoritmů. Předností také je vykreslení výsledné sociální sítě bez překrytí hran, v našem případě jsou použity algoritmy Kamada - Kawai a Force Scan Algorithm.

6.1 Protokol událostí ve formě XML souboru

Na začátku celého problému je nutné mít data, ze kterých chceme vytvořit danou sociální síť k následnému rozboru a nalezení vztahů mezi jednotlivými osobami. My předpokládáme, že data z byznys procesu čili protokol událostí, bude uložen ve formátu XML souboru.

Data uložená v XML souboru mají tu výhodu, že jsou v daném souboru zapsána podle dané struktury. Jelikož tuto strukturu uložení známe, je pro nás jednoduché se v daném souboru pohybovat a extrahovat námi potřebná data. Struktura daného souboru je znázorněna ve výpisu 1.

```
<?xml version="1.0" encoding="WINDOWS-1250"?>
<event_log>
  <event>
    <case>1</case>
    <activity>A</activity>
    <performer>John</performer>
  </event>
```

```

<event>
  <case>2</case>
  <activity>A</activity>
  <performer>Pete</performer>
</event>
...
</event.log>

```

Výpis 1: Ukázka struktury XML souboru

Pokud se blíže podíváme na danou ukázkou struktury XML souboru zobrazeného ve výpisu 1, vidíme, že na prvním řádku se dozvídáme o verzi specifikace, podle které je dokument napsán a nastaveném kódování daného souboru. Tento řádek představuje tzn. XML prolog. Poté následuje již samotná struktura, která obsahuje námi požadovaná data. Dokument je strukturovaně rozdělen pomocí elementů, které obsahují buď další elementy či samotná data. V této struktuře můžeme podle názvu elementu logicky usoudit, jaká data daný element obsahuje. Při bližším zaměření na strukturu vidíme, že element `<event.log>` obsahuje všechny další elementy a jedná se tedy o kořenový prvek. Při zanoření dále do struktury dostáváme element `<event>`, který obsahuje data jedné události v protokolu událostí. A dalším zanořením již dostáváme elementy `<case>`, `<activity>` a `<performer>`, které obsahují již samotná data, přesněji data o případu, data o aktivitě a data o uživateli v dané události.[9]

6.2 Zpracování dat

Jelikož máme soubor se strukturovanými daty, můžeme tyto data dále zpracovávat. Nejprve si ale musíme zvolit způsob, jak se na data budeme dívat, respektive jakou metrikou je budeme zpracovávat. V našem případě jsme zvolili **metriku založenou na společných případech**. Tato metrika je popsána v kapitole 4, my si ji však pro připomenutí vysvětlíme ještě jednou.

6.2.1 Metriky založené na společných případech

Metriky založené na společných případech, jak je z názvu patrné, se zaměřují na jednotlivé případy, ve kterých dané osoby vykonávají aktivity. Metrika jednoduše počítá, jak často dvě osoby vykonávají aktivity pro stejný případ. Pokud jednotlivci spolupracují na mnoha případech, budou mít silnější vztah než jednotlivci zřídka pracující společně.

Pokud se podíváme na implementaci tohoto problému, je potřeba k jednotlivým osobám, které se vyskytují v protokolu událostí, nalézt a uložit všechny případy, ve kterých provádějí aktivity. Následně můžeme hledat další osoby, které v daných případech rovněž provádějí operace a tímto způsobem nalézt všechny osoby, které pracují ve společných případech.

Prvním problémem je získání jednotlivých osob z protokolu událostí. Jelikož se osoby v protokolu událostí mohou vyskytovat několikrát, je důležité, abychom některou osobu

nezapočetli do výčtu vícekrát. Pro získání osob z protokolu událostí máme vytvořenou funkci `GetPerformers()`, která vrací pole obsahující seznam osob. Kód této funkce můžeme vidět ve výpisu 2.

```
public static string[] GetPerformers(string file)
{
    XDocument doc = XDocument.Load(file);

    var person = from l in doc.Descendants("performer")
                 select (string)l.Value;

    return person.Distinct().ToArray();
}
```

Výpis 2: Funkce pro získání všech osob z protokolu událostí

Další potřebnou funkcí je `GetUniqueCases()`, která vrací konkrétnímu uživateli respektive nastavuje potřebnou proměnnou instanci třídy `Node` pole případů, ve kterých daná osoba provádí aktivity. Opět je potřeba vracet jen výčet, který neobsahuje duplicity, to je vyřešeno metodou `Distinct()`.

```
public int GetUniqueCasesCount(string file)
{
    try
    {
        XDocument xdoc = XDocument.Load(file);

        var cases = (from ca in xdoc.Descendants("event")
                     where (string)ca.Element("performer") == mNode.mName
                     select (string)ca.Element("case").Value);

        mNode.mUniqueCases = cases.Distinct().Count();
        return 1;
    }
    catch (Exception e)
    {
        Console.WriteLine(e.Message);
    }

    return 0;
}
```

Výpis 3: Funkce pro získání všech případů dané osoby

Další funkce, která slouží pro získání potřebných dat z protokolu událostí je funkce `GetUniqueCasesCount()`, ta získá počet unikátních případů uživatele.

Z protokolů událostí dokážeme tedy dostat data v takové formě, kdy je můžeme opět používat a zpracovávat. Abychom dokázali určit vztahy mezi jednotlivými osobami, je

v programu nadefinované dvourozměrné pole o velikosti celkového počtu osob zúčastněných v daném byznys procesu. Pokud tedy budeme mít celkem patnáct osob, dvourozměrné pole bude mít rozměry `relationshipArray=[15, 15]`, kde každý řádek představuje jednu osobu, stejně jako každý sloupec v této matici. Hodnoty v této matici mohou nabývat určité kladné hodnoty, která představuje počet případů, které dané dvě osoby mají společné. Tato hodnota tedy představuje také sílu vztahu, čím vyšší hodnota, tím silnější vztah mezi dvěma uzly. Pokud hodnota v matici je rovna nule, znamená to, že osoby nepracují ani v jednom společném případě, nemají mezi sebou žádný vztah a tyto uzly nebudou ve výsledné sociální síti mezi sebou propojeny.

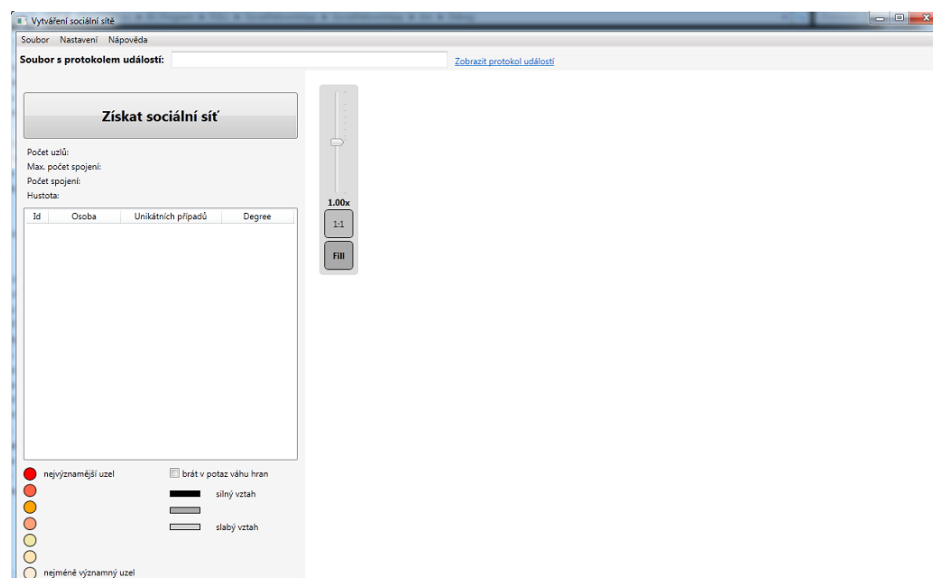
Mimo jiné si samozřejmě ukládáme do proměnných i další potřebná data a výpočty, které dále využíváme v programu. Ty však nejsou tak podstatné či zajímavé, abychom je zde všechny uváděli.

6.3 Zobrazení získaných výsledků

V této podkapitole si popíšeme nejen uživatelskou aplikaci, ale také zobrazení výsledků experimentů a jejich následné analýzy.

6.3.1 Uživatelská aplikace

Na obrázku 9 je znázorněn vzhled uživatelské aplikace, která slouží pro zobrazení výsledků z extrakce dat z protokolu událostí a jejich následné zpracování do sociogramu respektive sociální sítě a uvedení dalších neméně důležitých faktů o dané sociální síti a osob v ní.

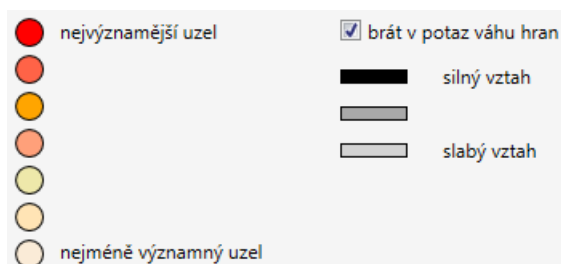


Obrázek 9: Uživatelská aplikace

Jak je vidět jedná se o lehce pochopitelné rozhraní, určené převážně pro experimenty. V pravé části se nachází grafické plátno, ve kterém se zobrazuje výsledná sociální síť. V horní části aplikace se nachází řádek pro zobrazení cesty k XML souboru, tedy k protokolu událostí. Dále lze v levé části rozpoznat tabulku, která slouží pro zobrazení jednotlivých uzlů v síti a další informace k těmto uzlům, jako je id daného uzlu, jméno osoby představující daný uzel, počet unikátních případů, ve kterých je daná osoba angažována a hodnotu degree (počet vztahů daného uzlu v síti k ostatním uzlům).

Dále aplikace počítá nejen celkový, ale i skutečný počet spojení mezi uzly v síti a vypočítává následnou hustotu výsledné sítě.

Výsledná zobrazená sociální síť je tvořena uzly, které mají v síti různou důležitost, která závisí na počtu spojení s ostatními uzly v této síti. Důležitost každého uzlu je snadno pochopitelná podle velikosti a barvy konkrétního vykresleného uzlu. Čím je uzel větší a barva ostřejší, tím je uzel v síti důležitější. Podobně je to také se spojeními mezi uzly, kde nejtmavší spojení představuje hranu s velkou váhou a naopak, světlý uzel hranu s malou váhou. V aplikaci je na výběr, zda chceme výslednou sociální síť vykreslit s ohledem na váhu jednotlivých hran či nikoliv. Jak velikost uzlu, tak barva hrany jsou v programu vypočítávány poměrně. Toto vše je v krátkosti vysvětleno v levé dolní části aplikace, viz obrázek 10.

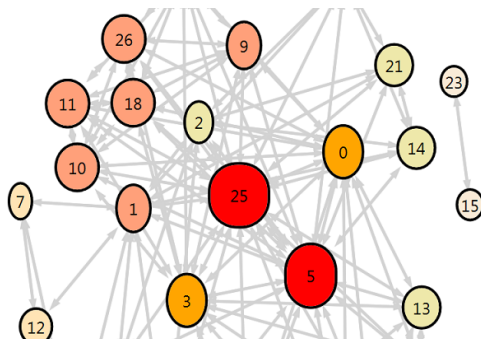


Obrázek 10: Vysvětlivky pro uzly a hrany

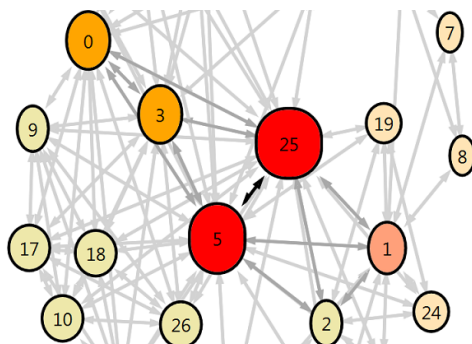
Pokud chceme vykreslit sociální síť, je prvně potřeba vybrat soubor s daty, která se mají zpracovat. V hlavní nabídce vybereme „Soubor“ a poté položku „Otevřít“, díky které se nám otevře pomocné okno pro nalezení XML souboru s protokolem událostí. Jakmile vybereme a potvrdíme vybraný soubor, v aplikaci se nám zobrazí cesta k tomuto souboru a pokud chceme, lze také samotný protokol událostí zobrazit. Po načtení vybraného souboru a kliknutí na tlačítko „Získat sociální síť“ se daný XML soubor zpracuje a vykreslí se sociální síť.

Na obrázku 11 můžeme vidět sociální síť, ve které se nebere zřetel na váhu jednotlivých hran, a tedy všechny hrany mají stejnou barvu. Naopak na obrázku 12 bereme v potaz váhu jednotlivých hran, a proto jsou některé hrany tmavší než ostatní. Všimněme si, že se hodnoty ve sloupci „Degree“ mění v závislosti na tom, zda máme či nemáme vybranou položku „brát v potaz váhu hran“. Je to z toho důvodu, že pokud položka

není vybrána, jde nám o to, zda uzly jsou či nejsou spojeny. Pokud máme však políčko zaškrtnuté, jde nám také o váhu hran mezi spojenými uzly.

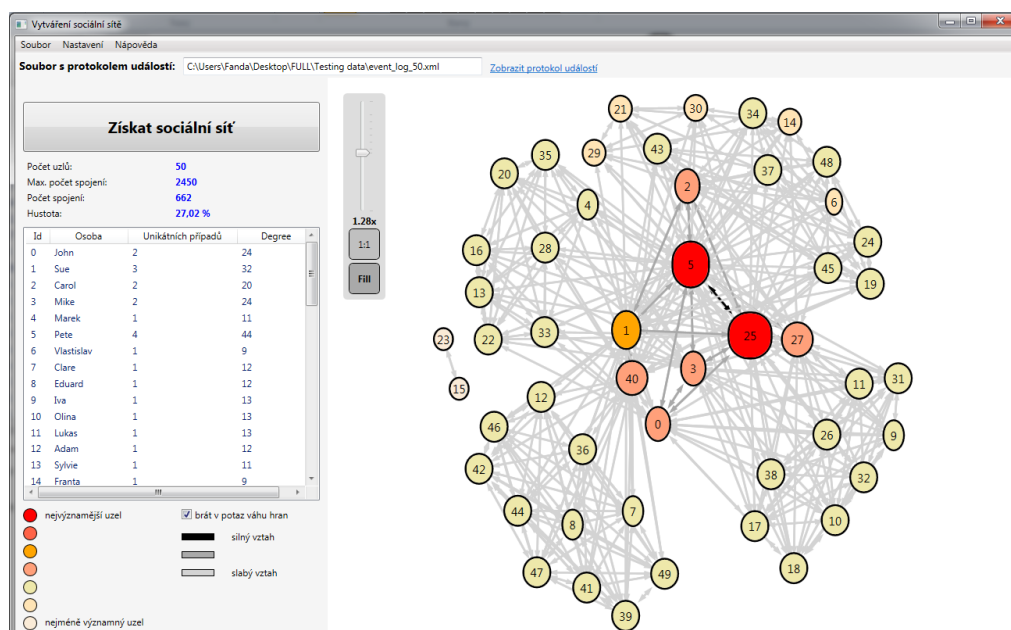


Obrázek 11: Detail sítě, kde jsou hrany vykresleny bez ohledu na jejich váhu



Obrázek 12: Detail sítě, kde jsou hrany vykresleny s ohledem na jejich váhu

Aplikace, ve které byl vybrán soubor s daty a kde byla tato data zpracována, je zobrazena na obrázku 13. V grafickém plátně je tedy vykreslená sociální síť založená na společných případech, kde jednotlivé uzly mají velikost v závislosti na tom, kolik z nich vychází hran (tedy s kolika dalšími uzly tvoří vztahy). V levé části aplikace je uvedená tabulka obsahující názvy všech osob zúčastněných v daném byznys procesu, jejich id, hodnotu Degree a počet unikátních případů, ve kterých jsou angažováni.

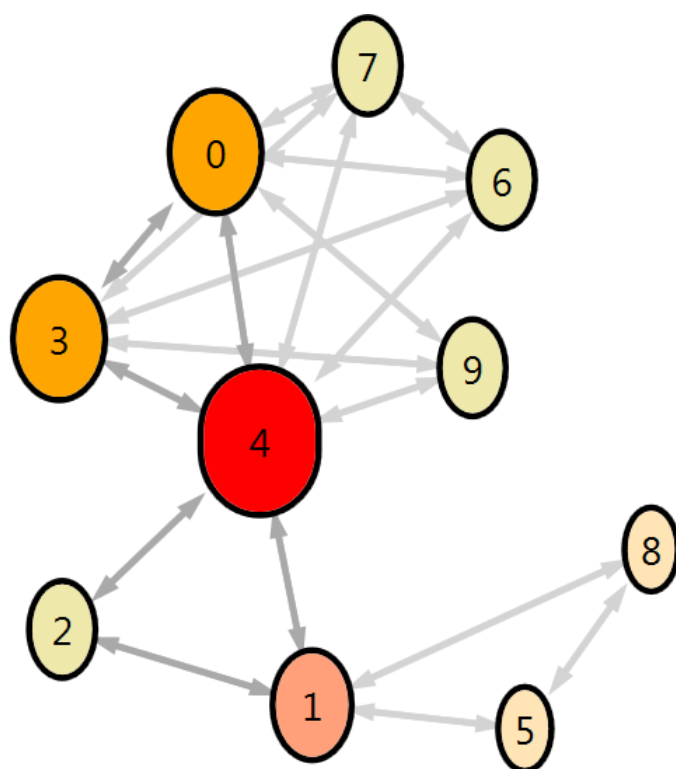


Obrázek 13: Zobrazení dat v uživatelské aplikaci

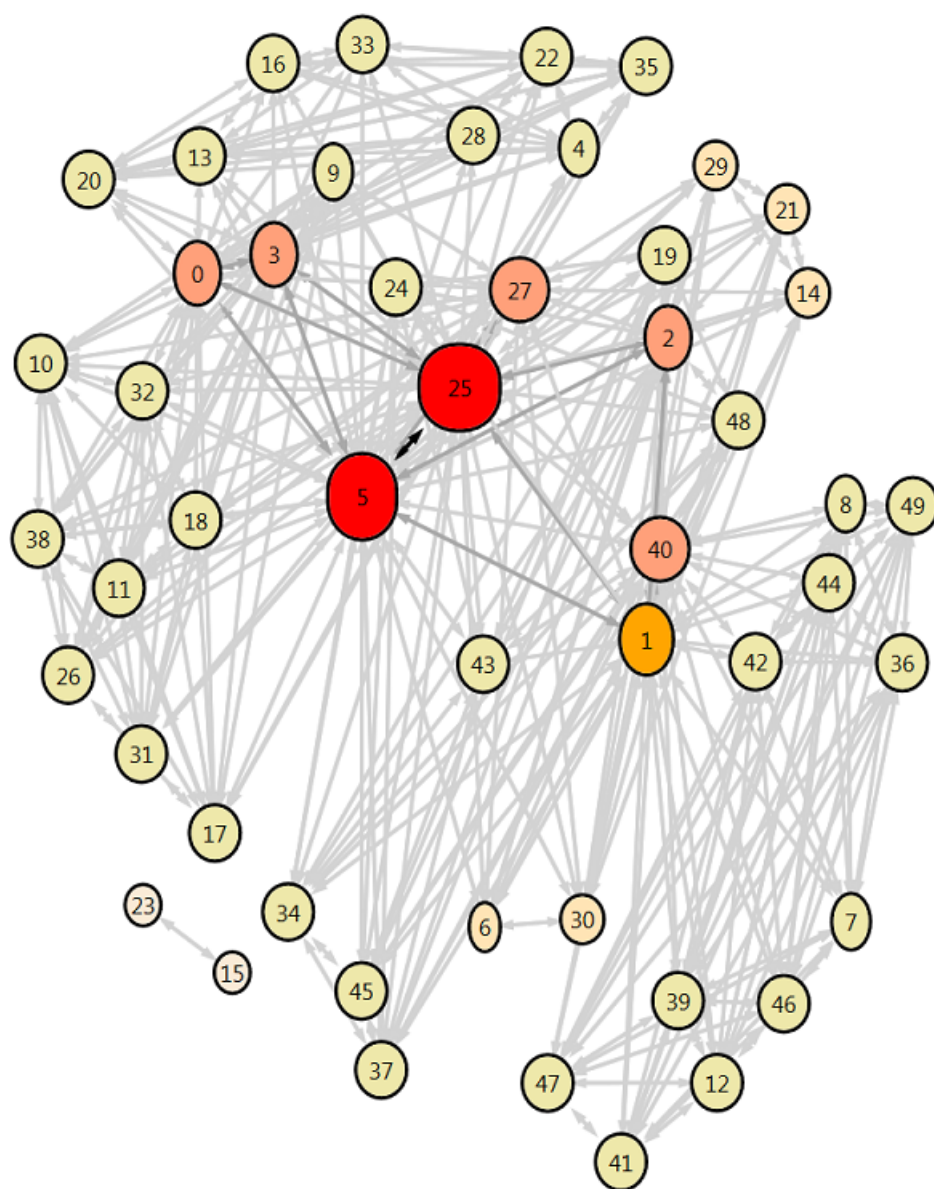
6.4 Popis experimentů

Experimenty probíhaly formou různě složitých a rozsáhlých protokolů událostí, které vytvořená uživatelská aplikace zpracovávala. S velmi malým počtem událostí v protokolu událostí a malým počtem osob se samozřejmě výpočty a vykreslení sociální sítě zpracují rychleji než s vzrůstajícím počtem dat. Bylo využito šest různých protokolů událostí, každý s jiným počtem zaznamenaných událostí i osob. Pro ukázkou se můžeme podívat na vykreslenou sociální síť z uživatelské aplikace, kde je angažováno deset osob a porovnat ji se sociální sítí, ve které je zapojeno osob padesát.

Z obrázku 14 a obrázku 15 můžeme vidět, že čím větší je počet uzlů a vztahů v sociální síti, tím se síť stává složitější a zpracování takovéto sítě se stává náročnější a čas pro zpracování se zvyšuje.



Obrázek 14: Vytvořená sociální síť s deseti uzly



Obrázek 15: Vytvořená sociální síť s padesáti uzly

7 Závěr

Při zpracovávání této bakalářské práce jsem získal velké množství znalostí nejen o formě protokolu událostí a získávání dat z něj, ale také převážně o využití těchto dat k dalšímu rozboru a analýze. Jak využít tyto data k vytvoření sociální sítě pomocí různých metrik, které mohou být k tomuto účelu použity. Následně jsem jednu z takto popsaných metrik k získání sociální sítě použil a implementoval. Důležitým bodem bylo také pochopení zobrazení výsledné sociální sítě, kdy opět existuje několik možností, jak lze síť zobrazit. Jelikož se jedná o graf, bylo nutné znát pojmy týkající se problematiky grafů. Samotná kapitola pojednávala o analýze takto vzniklé sociální sítě, kdy opět existuje nepřeberné množství metod a jejich obměn. Podobně jako u metrik sloužící k získání sociální sítě, jsem některé metody využitelné na analýzu popsal, následně jednu vybral a implementoval. Ve finální části tohoto dokumentu je popis implementace a uživatelské aplikace, kterou je možné použít na zobrazení sociální sítě z dat uložených v XML souboru.

Tato bakalářské práce zdaleka neobsahuje všechny možné postupy a metriky, které souvisejí s extrakcí sociálních sítí z byznys procesů. Existuje mnoho možností dalšího rozvoje této bakalářské práce. Ať už je to nastudování a popis dalších možných metrik a technik pro zpracování dat ze souboru či k využití analýzy výsledné sociální sítě, až po praktické využití a naprogramování těchto metrik a metod ve výsledné uživatelské aplikaci.

Pokud by byla příležitost, určitě bych se rád tomuto tématu v budoucnu dále věnoval, jelikož se jedná o velmi zajímavé téma skrývající ještě mnoho dalších možností pro prozkoumání a rozšíření.

Lukáš Panchártek

8 Reference

- [1] van der Aalst, W.M.P., Reijers, H.A., Song, M.: *Discovering Social Networks from Event Logs*, Computer Supported Cooperative Work , 2005
- [2] van der Aalst, H.A., Song, M.: *Mining Social Networks: Uncovering Interaction Patterns in Business Processes*, New York: Addison-Wesley Pub. Co., 2004.
- [3] Vondrák, Ivo: *Metody byznys modelování*, Ostrava, 2004.
- [4] *Business process* [online]. c2013, [cit. 2013-04-14]
http://en.wikipedia.org/wiki/Business_process
- [5] *Social Network* [online]. c2013, [cit. 2013-04-14]
http://en.wikipedia.org/wiki/Social_network
- [6] *Centrality* [online]. c2013, [cit. 2013-04-14]
<http://en.wikipedia.org/wiki/Centrality>
- [7] *Sociogram* [online]. c2013, [cit. 2013-04-14]
<http://cs.wikipedia.org/wiki/Sociogram>
- [8] *Social Network Analysis, A Brief Introduction* [online]. c2013, [cit. 2013-04-14]
<http://www.orgnet.com/sna.html>
- [9] *Základy XML* [online]. c2013, [cit. 2013-04-14]
<http://homel.vsb.cz/~sla10/educ/EPubl/XML-moznosti/output/ch04s01.html>
- [10] *Microsoft, Visual C#* [online]. c2013, [cit. 2013-04-14]
<http://msdn.microsoft.com>
- [11] Moody, James: *Centrality in Social Network* [online]. c2013, [cit. 2013-04-14]
www.soc.duke.edu/~jmoody77/s884/notes/class_centrality.ppt
- [12] *GraphSharp* [online]. c2013, [cit. 2013-07-10]
<http://graphsharp.codeplex.com/>

A Příloha na CD/DVD

CD, které je přílohou této bakalářské práce, obsahuje tento dokument v elektronické formě a program se vzorovými daty. Rozmístění jednotlivých částí je zobrazeno v následující tabulce.

\dokument	obsahuje tento dokument v elektronické podobě
\program	zdrojové kódy programu
\testovací_aplikace	spustitelná aplikace pro extrakci a analýzu sociálních sítí
\vzorova_data	xml soubory, obsahující protokoly událostí